Impact of scaffolding and question structure on the gender gap

Journal Item

Impact of scaffolding and question structure on the gender gap

Hillary Dawkins, Holly Hedgeland, and Sally Jordan
School of Physical Sciences, The Open University, Walton Hall, Milton Keynes MK7 6AA, United Kingdom
(Received 24 April 2017; published 13 September 2017)

We address previous hypotheses about possible factors influencing the gender gap in attainment in physics. Specifically, previous studies claim that scaffolding may preferentially benefit female students, and we present some alternative conclusions surrounding this hypothesis. By taking both student attainment level and the degree of question scaffolding into account, we identify questions that exhibit real bias in favor of male students. We find that both multidimensional context and use of diagrams are common elements of such questions.

DOI: 10.1103/PhysRevPhysEducRes.13.020117

I. INTRODUCTION

The gender gap in attainment in physics is consistent and well documented. Across institutions, male students outperform their female counterparts in terms of undergraduate course performance [1–3], as well as outcome on subject specific concept inventories (FCI [1,4–7], BEMA [8–10], and CSEM [10,11]). At the UK Open University (OU), we observe a significant difference in attainment on the second level physics modules in favor of males, and, furthermore, this gap is persistent across multiple years of instruction.

While the existence of a real and significant gap is well established, the contributing factors are less well understood (see Ref. [12] for a review of 17 studies). Possible factors include background and preparation, of which many possible measures exist. Previous studies identify concept inventory pretest scores [9,13], SAT math scores [7,9,13], ACT math scores [9,13], and prerequisite course grades [9] to vary significantly by gender. Sociocultural factors may also play a role, for example, self-efficacy and CLASS scores [14] (a measure of learning attitudes about science). Finally, there is the issue of question construction including type of question (constructed response, multiple choice, or other selected response), presentation (graphs, diagrams, words), and male-biased context (references to sports and cannons). Here we focus on identifying factors from the final category of question structure, as these are the most readily modified.

A recent study from the University of Cambridge [15] observes an interesting dependence on question structure in the form of scaffolding. Scaffolding refers to the degree to which a question guides the student through the problem-solving process. Previous studies support the use of scaffolding in aiding students' learning and conceptual understanding in physics [16–18]. However, Ref. [15] is the first study to our knowledge to identify a dependence on gender. It is therefore important to verify these findings across institutions and student populations prior to taking action towards any instructional reform.

In light of the large and diverse student population of the Open University, we find ourselves well situated to address these issues. The goals of the present study are to

(1) Identify elements of question structure which may be disadvantaging female students.
(2) Test the scaffolding hypothesis as a potential solution.

Taking student ability (as measured by overall attainment levels) and question difficulty into account, we identify questions that pose significant male bias and those which do not. We discuss our findings in the context of current literature on the subject. Furthermore, we challenge the conclusions presented in Ref. [15], and offer some alternative conclusions.

II. CONTEXT

The present study examines gender differences in attainment observed in the second level (FHEQ Level 5) physics modules at the Open University. We first spend some time reviewing the structure of the Open University, the modules in question, and the student population.

The Open University approaches higher education in a nontraditional way in that there are no admission requirements, and modules are completed at a distance with substantial online elements. Students select and complete modules, according to their needs, to make up a degree comprised of 360 credits if desired. Students are attracted to the open concept for a variety of reasons including flexibility, part-time options, returning to study later in life, and completing second degrees. We therefore expect that the student population is demographically diverse. Despite differences in the student population, similar trends in attainment gaps have been identified as at other institutions.

Of particular interest is a large gap in attainment at the second level, the first level at which physics is taught as a separate module, which does not exist at lower or higher levels.

The 60-credit second level physics modules (previously S207, now S217) include mechanics, thermodynamics, electricity and magnetism, quantum physics, and nuclear physics at an introductory to intermediate level. Although prerequisites are not enforced, it is expected that students will have completed the introductory level one science module, from which they will have gained some familiarity with some of these topics as well as appropriate mathematical preparation. The module population comprises a mixture of students intending to take further physics modules and those intending to take further science or mathematics modules outside of physics.

Throughout the module, students complete interactive computer-marked assignments (iCMAs), which are short problems requiring numeric open responses or selected responses, in addition to tutor-marked assignments. Students receive feedback on their iCMA answers and are permitted to retry questions as many times as desired. The module ends with an exam that contains, among other components, long-answer open response questions. In this study, we analyze iCMA questions to identify any gender bias, and look at exam long-answer questions to address the scaffolding hypothesis. Data were collected over four recent presentations of the module: S207 in 2012–2014 and S217 in 2015. The total number of students completing the module in this time period was 1727: 1335 (77%) males and 392 (22%) females.

III. IDENTIFYING BIAS

A. The Mantel-Haenszel method

The Mantel-Haenszel method is a statistical technique used to identify differences between groups using a stratified data set [19]. The idea is that possible confounding variables will be captured by the stratification.

In this case, we wish to detect iCMA questions that exhibit significant male bias while accounting for student ability and question difficulty. Therefore, we take our two groups to be male and female students, and students are stratified according to ability as measured by their overall performance on iCMA questions. Table I shows a cross table representing the number of students in each group answering an item correctly at the ith stratum. For each item, we calculate the odds ratio (ratio of success probabilities between groups) of the ith stratum as $m_i^1 f_i^0 / m_i^0 f_i^1$. A weighted average across all strata then provides the overall odds ratio for a particular question, referred to as the Mantel-Haenszel alpha:

$$\alpha_{\mathrm{MH}} = \frac{\sum_i m_i^1 f_i^0 / N_i}{\sum_i m_i^0 f_i^1 / N_i}. \tag{1}$$

TABLE I. A cross table of the ith stratum depicting the number of students in each group (male and female) to get a particular iCMA question correct or incorrect on the first attempt. The total number of students in the ith stratum is $N_i = m_i^1 + m_i^0 + f_i^1 + f_i^0$.

	Correct (1)	Incorrect (0)
Male	m_i^1	m_i^0
Female	f_i^1	f_i^0

For comparison, this is often converted to a logarithmic scale as

$$\alpha^*_{\mathrm{MH}} = -2.35 \ln\left(\frac{\sum_i m_i^1 f_i^0 / N_i}{\sum_i m_i^0 f_i^1 / N_i}\right). \tag{2}$$

The sign and magnitude of α^*_{MH} indicate the direction and strength of bias within a question. Negative values indicate a bias in favor of males, meaning that male students have a greater probability of answering this question correctly compared to female students of equal ability. Likewise, positive values indicate a bias in favor of females. The absolute value of α^*_{MH} indicates the strength of the bias, and is deemed to be significant if $|\alpha^*_{\mathrm{MH}}| \geq 1$ [20].

As a second assurance of significance, each α_{MH} value is tested using a chi-squared distribution. In this case, the null hypothesis is that the odds ratio is equal to 1 at each stratum, and the alternative hypothesis is that at least 1 odds ratio is different from unity [19]. In this study, we flag questions as having significant bias if both conditions (i) $|\alpha^*_{\mathrm{MH}}| \geq 1$ and (ii) $p \leq 0.05$ are satisfied.

B. Analysis and results

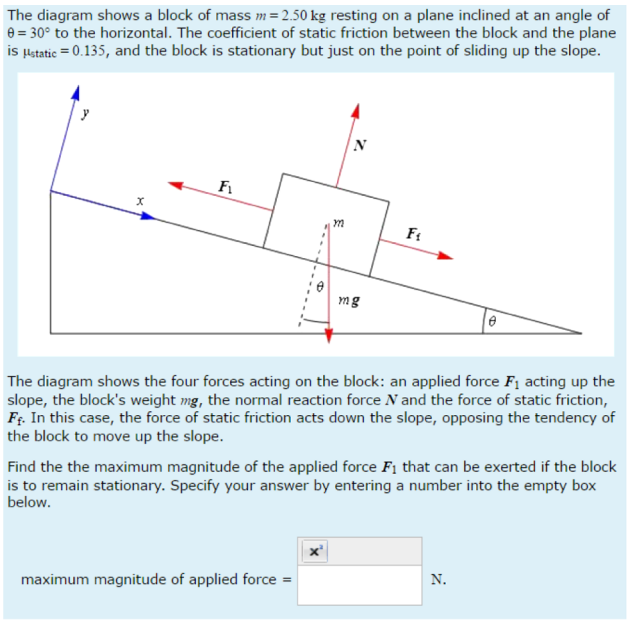
The Mantel-Haenszel method was applied to all 56 iCMA questions that were presented to students in the four presentations of the S207/S217 modules included in this study. iCMA items evenly span the varied course content, and are meant to be short problems involving minimal calculations. The analysis flags 3 questions of significant bias, all in favor of male students. Further, 2 questions were noted to be of interest having significant p values and insignificant $|\alpha^*_{\mathrm{MH}}|$ values, but in favor of female students. These items were included for being the only questions of some significance with female bias. Table II shows the α^*_{MH} values with significance levels for each question of interest. Questions are labeled as M_1, M_2, M_3 (those having male

TABLE II. Strength of bias (α^*_{MH}) and significance (p) values for iCMA questions of interest.

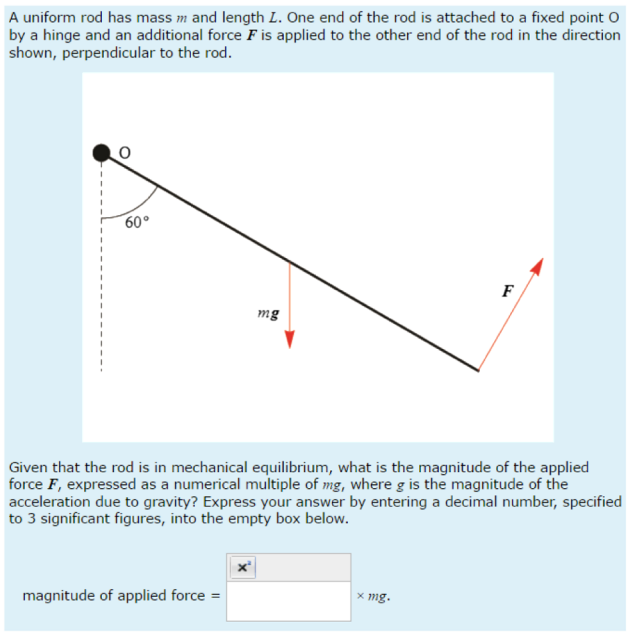
	M_1	M_2	M_3	F_1	F_2
α^*_{MH}	−2.9	−3.2	−3.3	0.14	0.39
p	0.045	0.013	0.032	0.016	0.037

advantage) and F_1, F_2 (those having female advantage). We note that M_1, M_2, and M_3 all display very strong levels of bias with $|\alpha^*_{\mathrm{MH}}|$ values well above the threshold.

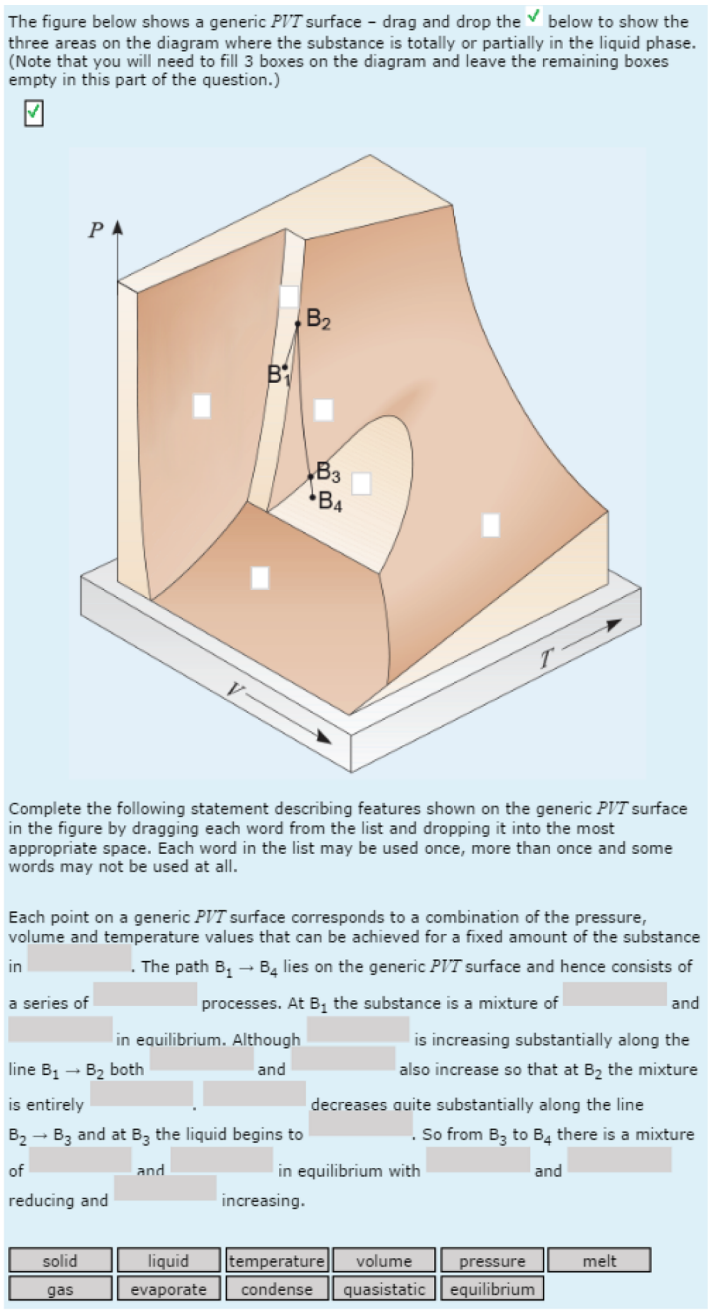
Figure 1 shows the items displaying male bias. Notably, all questions require interpreting a diagram of more than one dimension, which we find to be consistent with current literature. Wilson *et al.* [21] studied the impact of question structure on the gender gap along five broad dimensions: content, process required, difficulty, presentation, and context. They observed large gender gaps in favor of males for questions which involved the process of interpreting a diagram, which presented the question using a significant diagram, and which involved more than one spatial dimension. Studies that aim to identify gender gaps on FCI questions have observed the largest disparities on items 6 (path of ball leaving a channel), 12 (path of cannonball fired off a cliff) [22], 14 (path of object released from an airplane), and 23 (path of a rocket after thrust is turned off) [1]. Clearly, all of these items involve predicting motion in two dimensions, and all are presented using a diagram. The



(a) M_1: Inclined plane



(b) M_2: Torque



(c) M_3: Generic PVT surface

FIG. 1. iCMA items displaying bias in favor of male students.

observed gender gap on projectile-motion-like items is sometimes ascribed to male-biased context [23]. However, attempts to reword FCI items in a more traditionally female context have failed to improve female performance [24]. In light of this discussion we find the inclusion of item M_3 particularly interesting. The content is thermodynamics, far removed from kinematics or predicting motion. The context is certainly not experienced or male biased, and yet a large and significant gap is observed. The only identifiable common trait among all items is the need to interpret a multidimensional diagram. We note that some items which were not flagged as having significant male bias do contain multidimensional diagrams. These included two-dimensional graphs, which we hypothesize were not challenging to students due to mathematical familiarity, and circuit diagrams, for which spatial interpretation is not very relevant. Furthermore, we emphasize that our analysis does not claim that unflagged items exhibit no bias, only that flagged items exhibit strong bias.

Figure 2 shows the items displaying female bias. As previously stated, these questions have significant p values but do not have significant $|\alpha^*_{\mathrm{MH}}|$ values, implying that the bias is small. Nonetheless, these items are of interest as the only female-biased questions of some significance. Both items involve careful reading, a task suggested to have a female advantage [25]. Interestingly, item F_2 is on the subject of predicting motion. This observation further supports the idea that male bias arises from the need to interpret a diagram or multidimensional context, rather than content related to predicting motion.

IV. SCAFFOLDING

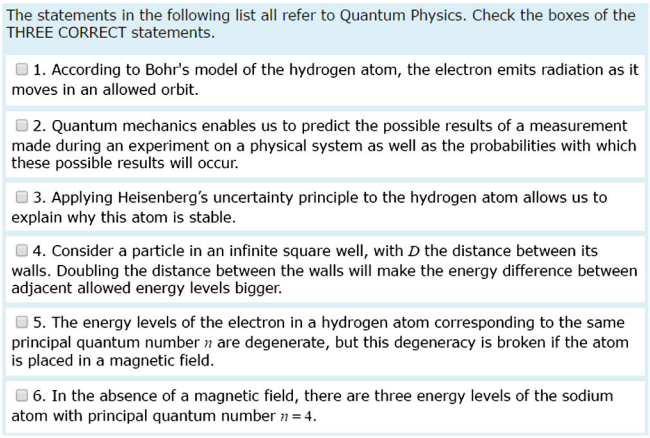
A. Scaffolding definition

Scaffolding is broadly defined to have occurred when an expert or more knowledgeable person helps a learner to accomplish tasks that would otherwise be unattainable [26]. A traditional example would be a teacher providing strategic guidance and feedback while a student completes a problem. In more recent years this definition has evolved to include interactive computer-assisted learning, as well as peer instruction and similar socialized learning environments [27].

Because of widespread usage of the term "scaffolding" in multiple circumstances, it is important to carefully define the term in the context of physics education research. In the present study, we consider scaffolding only as it may be applied to written exam questions. We define six general ways in which scaffolding can occur (elements), and further provide specific instances of each that are likely to be encountered in physics problems. Table III shows a complete itemization of the elements. Many elements are adapted from the guidelines outlined in Ref. [28], which combines theoretical foundations with prior work to define a common framework for scaffolding within computer-assisted assignments. The element of conceptual prompting is motivated by Ding *et al.* [16]. There it was shown that students will successfully apply physics concepts to problems if they are prompted to identify the concept immediately beforehand. Taken together, the elements listed in Table III define what is meant by scaffolding within this study.

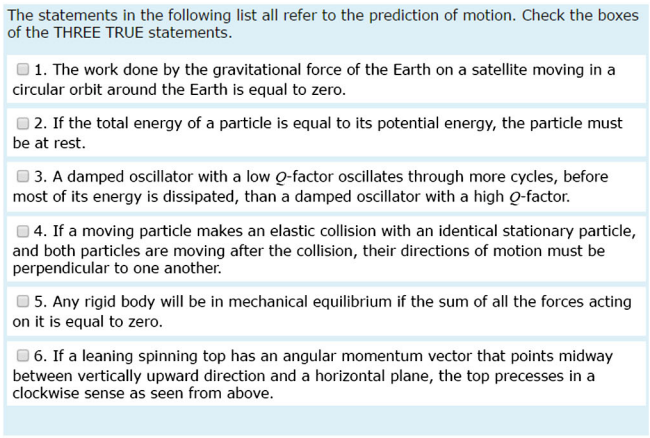
B. Gains by gender

In a study on question structure and its impact on the gender gap, Gibson *et al.* [15] administered two versions of an exam. One exam used highly scaffolded questions, and one used traditional exam style questions. Between the low and high scaffolding versions, female students achieved a gain in exam score of 13.4% while male students achieved a gain of 9.0%. The study therefore concludes that scaffolding benefits all students, but that female students benefit preferentially. We observe no such preferential treatment, and argue that other factors may be at play.


The statements in the following list all refer to Quantum Physics. Check the boxes of the THREE CORRECT statements.

1. According to Bohr's model of the hydrogen atom, the electron emits radiation as it moves in an allowed orbit.
2. Quantum mechanics enables us to predict the possible results of a measurement made during an experiment on a physical system as well as the probabilities with which these possible results will occur.
3. Applying Heisenberg's uncertainty principle to the hydrogen atom allows us to explain why this atom is stable.
4. Consider a particle in an infinite square well, with D the distance between its walls. Doubling the distance between the walls will make the energy difference between adjacent allowed energy levels bigger.
5. The energy levels of the electron in a hydrogen atom corresponding to the same principal quantum number n are degenerate, but this degeneracy is broken if the atom is placed in a magnetic field.
6. In the absence of a magnetic field, there are three energy levels of the sodium atom with principal quantum number $n = 4$.

(a) F_1: Quantum physics


The statements in the following list all refer to the prediction of motion. Check the boxes of the THREE TRUE statements.

1. The work done by the gravitational force of the Earth on a satellite moving in a circular orbit around the Earth is equal to zero.
2. If the total energy of a particle is equal to its potential energy, the particle must be at rest.
3. A damped oscillator with a low Q-factor oscillates through more cycles, before most of its energy is dissipated, than a damped oscillator with a high Q-factor.
4. If a moving particle makes an elastic collision with an identical stationary particle, and both particles are moving after the collision, their directions of motion must be perpendicular to one another.
5. Any rigid body will be in mechanical equilibrium if the sum of all the forces acting on it is equal to zero.
6. If a leaning spinning top has an angular momentum vector that points midway between vertically upward direction and a horizontal plane, the top precesses in a clockwise sense as seen from above.

(b) F_2: Predicting motion

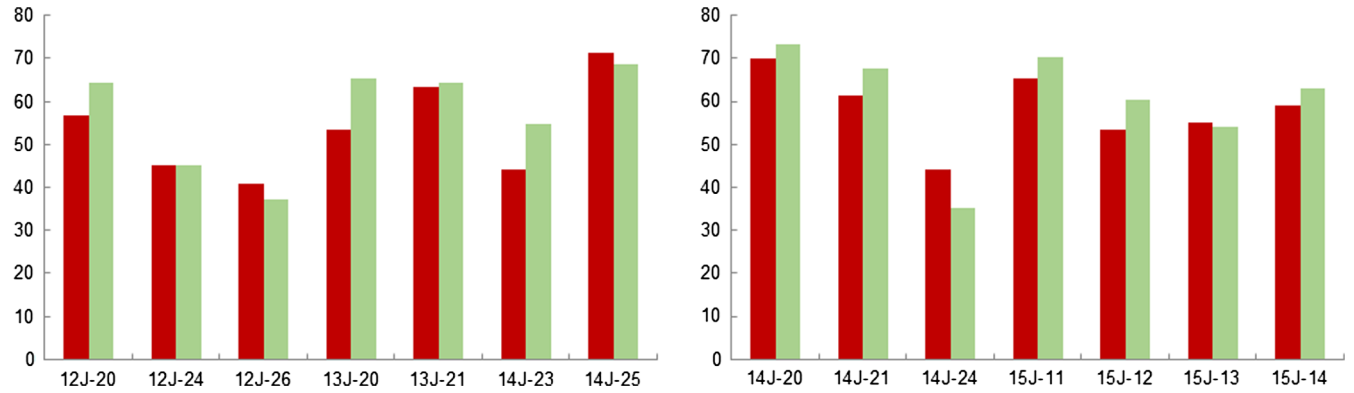
FIG. 2. iCMA questions displaying bias in favor of female students.

TABLE III. The 6 elements of scaffolding (bold), with itemized examples of how each element is likely to appear in written physics problems.

use of representations and language to bridge expert-novice understanding
1. technical words are described in everyday language
2. mathematical symbols are explained in words
3. a diagram is used to give meaning to technical words or symbols
reduction of cognitive overhead
4. includes a math (or other background) reminder
5. somehow automates a routine task (e.g., unit conversions given, constants given that could have been looked up)
6. no penalty for missing sig figs, wrong unit, wrong numeric value, or other nonsalient component of the question
7. provides a diagram or graph that the student could have constructed with the available information
insertion of expert knowledge
8. expert directed focus is used (e.g., key information is highlighted using bold or italicized text)
9. explicitly instructs student to make an expert assumption (e.g., "you may ignore air resistance")
10. the student is warned of a common mistake or relevant misconception
ordered task decomposition (provide structure for complex tasks)
11. each part of the question contains only one expected output (numeric or otherwise)
12. an output (numeric or otherwise) is required in subsequent work
13. marks are awarded for interpreting outputs (no further calculation required)
14. question has a wide mark distribution (each part is worth less than 50% of the total awarded marks)
conceptual prompting
15. asks student to define or explain an equation that they should use
16. asks student to identify a concept that they should make use of
17. asks student to draw a diagram before beginning the problem
reduction of degrees of freedom
18. gives student the appropriate equation to use
19. prompts at how the question is expected to be solved (e.g., "using the principle of conversation of energy …")
20. explicitly instructs student on how to begin a task

The present study includes all long-answer exam questions administered to students during four presentations of S2017/S217 (2012–2015). Using the elements of scaffolding and individual items as a scoring system, all exam questions were assigned a "scaffolding score." A question's score is equal to the number of items it contains, where multiple occurrences of items are counted. Questions displaying two or fewer items were labeled as low scaffolding, and questions with 7 or more items were labeled as high scaffolding. The values of 2 and 7 were chosen to ensure a large difference in scaffolding between groups. All questions belonging to either group can be found in the Supplemental Material [29]. Applying the same scoring system to the exam questions used in the study of Gibson *et al.*, we find that questions labeled as high scaffolding there do score highly by our definition. Referring to the elements of scaffolding, it is clear that Gibson *et al.* have employed ordered task decomposition extensively, as well as conceptual prompting (in the form of asking students to identify equations and draw diagrams before beginning), in their design of highly scaffolded questions. We therefore conclude that Ref. [15] and the present study use comparable definitions of scaffolding.

Figure 3 shows the performance of students on each question by gender, and Table IV shows the average performance as well as gains provided by increased scaffolding. The average gain is 6.6% for female students, and 5.2% for male students. Although not as clear, the data do at first glance seem to support the conclusions of Ref. [15]. Male students outperform female students on the low scaffolding questions by 2.9% ($p = 0.087$), and by only 1.4% on the high scaffolding questions ($p = 0.42$). However, neither result is statistically significant, and we should also consider how scaffolding benefits students performing at different levels. Intuitively, we expect that scaffolding cannot greatly benefit the highest achieving students (who likely know the information and do not have much room to improve) or the lowest achieving students (who are too unprepared for scaffolding to provide a use). Students completing module S207 and S217 are assigned a level (1–4) based on overall performance on the module (1 being the highest level of achievement). Table V shows the average score of students on the low and high scaffolding questions by level, as well as the number of male and female students in each level. As expected, scaffolding provides the greatest gains to the intermediate students. Performing a weighted average of gains across level by the number of female and male students in each level can give us an idea of the expected gains by gender. Doing this, we estimate expected gains of 6.2% for female



(a) Low scaffolding questions (b) High scaffolding questions

FIG. 3. Average performance (percentage score) of female (red) and male (green) students on all exam questions belonging to the low (a) and high (b) scaffolding groups. Exam questions are labeled as they appear in the Supplemental Material [29].

students, and 5.8% for male students. The expected gain is higher for female students as a consequence of the fact that fewer female students achieve a level 1. The expected gains are not significantly different than the actual gains for either gender, and therefore we conclude that preferential female gain is simply an artifact of gain dependency on level.

C. Questions of interest

Although scaffolding does not appear to preferentially benefit female students in general, we note some particular questions of interest. Figure 4 shows one question from the low scaffolding group (L), and one question from the high scaffolding group (H). Both are two-dimensional projectile

TABLE IV. Average performance by gender on exam questions assigned to the low and high scaffolding groups. Difference represents the average difference in exam grade between genders. Gain represents the increase in average exam score as a result of increased scaffolding.

	Male	Female	Difference
Low scaffolding	65.4	62.5	2.9
High scaffolding	70.5	69.1	1.4
Gain	5.2	6.6	

TABLE V. Average performance by level on exam questions assigned to the low and high scaffolding groups. Gain represents the increase in average exam score as a result of increased scaffolding. N represents the total number of students achieving each level by gender.

Level	1	2	3	4
Low average	90.5	69.5	52.5	40.7
High average	90.8	76.5	61.2	46.1
Gain	0.33	7.0	8.7	5.4
N males	414	732	577	297
N females	85	248	151	71

An astronaut playing golf on the Moon hits a ball so that it is initially moving with a speed of $u = 8.00\ \mathrm{m\ s^{-1}}$ at an angle of $\theta = 30°$ to the horizontal. In the following, the magnitude of the acceleration due to gravity at the Moon's surface, g_M, is approximately $1.62\ \mathrm{m\ s^{-2}}$.

(a) Make sketches of the vertical components of displacement, s_y, and velocity, v_y, versus time, t. Label the sketches with appropriate equations for s_y and v_y.

(6 marks)

(b) Assuming the surface of the Moon is flat in the vicinity of the astronaut, calculate how far the ball travels.

(4 marks)

(a) L: Low scaffolding

An athlete competing in the hammer throw event swings a heavy metal ball on a wire around in a circle. The radius of the circle that the ball travels is $1.5\ \mathrm{m}$ and the ball takes $0.55\ \mathrm{s}$ to complete one revolution.

(a) Calculate the *magnitude* of the instantaneous velocity of the ball and state the *direction* of the velocity at any instant, relative to the circle.

(2 marks)

At the instant when the velocity of the ball is in a direction at $50°$ to the horizontal, moving upwards, the ball is released. At this instant, the ball is $2.0\ \mathrm{m}$ above the ground, as shown in Figure 2.

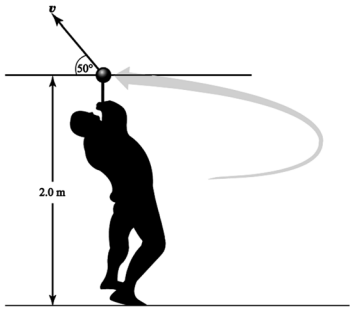


Figure 2 For use with Q20.

(b) What are the horizontal (x) and vertical (y) components of the ball's velocity at the instant when it is released?

(3 marks)

(c) What is the maximum height above the ground attained by the ball during its flight?

(5 marks)

(You may assume that the magnitude of the acceleration due to gravity is $9.8\ \mathrm{m\ s^{-2}}$ and you may ignore air resistance.)

(b) H: High scaffolding

FIG. 4. A pair of two-dimensional projectile motion problems of different scaffolding levels. Male students very significantly outperform female students on L, but performance is equal across gender on H.

TABLE VI. Average performance of students on questions of interest L and H by gender. The difference between male and female attainment is displayed along with the significance level (p values).

	Male	Female	Difference	Significance (p)
L	64.5	56.6	7.8	0.013
H	73.3	70.0	3.3	0.46

motion questions, but display significant performance differences. Table VI shows the average performance on each question, and the difference between genders with significance levels. Of all exam questions, L exhibits one of the most significant differences in performance between genders, and H shows no significant difference. The scaffolding gains are comparable to those observed in Ref. [15] (13.4% for females, 8.8% for males). We conclude that scaffolding may play a role in reducing the gender gap in specific types of problems that were previously identified to contain a male bias, namely, questions involving multidimensional context.

V. DISCUSSION AND CONCLUSIONS

In summary, we have identified elements of question structure that promote male bias, and further address the scaffolding hypothesis as a potential solution. We conclude that the level of scaffolding cannot sufficiently explain the gender gap.

We have used a Mantel-Haenszel stratified analysis to account for student ability, and find iCMA questions with significant performance differences between genders. By flagging only those questions that display significant bias in both measures ($|\alpha^*_{\mathrm{MH}}|$ and p), we have reduced the possibility of flagging false positives. We therefore conclude that the 3 flagged questions exhibit real and significant male bias. All questions involve interpreting a diagram, and all involve multidimensional context. Our findings are in agreement with Ref. [21], and similar studies on the FCI [1,22]. Because multidimensional diagrams appear most frequently in mechanics problems, previous studies may have incorrectly attributed male bias to mechanics content. Further investigation with more types of questions will be required to separate the variables of content and presentation.

Scaffolding has recently been argued to preferentially benefit female students [15], and therefore have the potential to aid in reducing the gender gap. The study of Gibson *et al.* uses a smaller number of students and less varied exam content than the present study to reach this conclusion. In a similar analysis, we do not observe a dependence on gender, and argue that any perceived dependence is actually due to student achievement level. The advantage of Ref. [15] is that exam questions were designed specifically to measure scaffolding gains, whereas the present study collected data from actual exam responses. Therefore, questions between the low and high scaffolding groups do not match onto each other exactly as in Ref. [15]. Future studies can make use of the elements of scaffolding to produce low and high scaffolding versions of the same question for use in experimental exams.

Even if scaffolding does not preferentially benefit female students in general, it may still play a role in reducing the gender gap. We make note of a pair of questions involving multidimensional context (2D projectile motion), for which the gap is reduced between low and high scaffolding versions. If male bias within a question can be reduced by increased scaffolding for novice students, then this provides a route to addressing gender gaps in attainment.

ACKNOWLEDGMENTS

The authors acknowledge financial support from eSTEeM, the OU centre for STEM pedagogy, as well as the cooperation of the S207 and S217 module teams, and useful discussions with Richard Jordan.

[1] J. Docktor and K. Heller, Gender differences in both Force Concept Inventory and Introductory Physics Performance, AIP Conf. Proc. **1604**, 15 (2008).
[2] A. Wee, B. Baaquie, and A. Huan, Gender differences in undergraduate physics examination performance and learning strategies in Singapore, Phys. Educ. **28**, 158 (1993).
[3] S. Andersson and A. Johansson, Gender gap or program gap? Students' negotiations of study practice in a course in electromagnetism, Phys. Rev. Phys. Educ. Res. **12**, 020112 (2016).
[4] S. Bates, R. Donnelly, C. MacPhee, D. Sands, M. Birch, and N. R. Walet, Gender differences in conceptual understanding of Newtonian mechanics: A UK cross-institution comparison, Eur. J. Phys. **34**, 421 (2013).
[5] E. Brewe, V. Sawtelle, L. H. Kramer, G. E. O'Brien, I. Rodriguez, and P. Pamelà, Toward equity through participation in Modeling Instruction in introductory university physics, Phys. Rev. ST Phys. Educ. Res. **6**, 010106 (2010).
[6] C. T. Richardson and B. W. O'Shea, Assessing gender differences in response system questions for an introductory physics course, Am. J. Phys. **81**, 231 (2013).
[7] M. Lorenzo, C. H. Crouch, and E. Mazur, Reducing the gender gap in the physics classroom, Am. J. Phys. **74**, 118 (2006).

[8] S. Lauer, J. Momsen, E. Offerdahl, M. Kryjevskaia, W. Christensen, and L. Montplaisir, Stereotyped: Investigating gender in introductory science courses, CBE Life Sci. Educ. **12**, 30 (2013).

[9] L. E. Kost-Smith, S. J. Pollock, and N. D. Finkelstein, Gender disparities in second-semester college physics: The incremental effects of a smog of bias, Phys. Rev. ST Phys. Educ. Res. **6**, 020112 (2010).

[10] S. J. Pollock, Comparing student learning with multiple research-based conceptual surveys: CSEM and BEMA, AIP Conf. Proc. **1604**, 171 (2008).

[11] P. B. Kohl and H. V. Kuo, Introductory physics gender gaps: Pre and post-Studio transition, AIP Conf. Proc. **1179**, 173 (2009).

[12] A. Madsen, S. B. McKagan, and E. C. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?, Phys. Rev. ST Phys. Educ. Res. **9**, 020121 (2013).

[13] L. Kost, S. Pollock, and N. Finkelstein, Characterizing the gender gap in introductory physics, Phys. Rev. ST Phys. Educ. Res. **5**, 010101 (2009).

[14] W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman, New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey, Phys. Rev. ST Phys. Educ. Res. **2**, 010101 (2006).

[15] V. Gibson, L. Jardine-Wright, and E. Bateman, An investigation into the impact of question structure on the performance of first year physics undergraduate students at the University of Cambridge, Eur. J. Phys. **36**, 045014 (2015).

[16] L. Ding, N. Reay, A. Lee, and L. Bao, Exploring the role of conceptual scaffolding in solving synthesis problems, Phys. Rev. ST Phys. Educ. Res. **7**, 020109 (2011).

[17] S. Lin and C. Singh, Effect of scaffolding on helping introductory physics students solve quantitative problems involving strong alternative conceptions, Phys. Rev. ST Phys. Educ. Res. **11**, 020105 (2015).

[18] C. Lindstrø m and M. D. Sharma, Teaching physics novices at university: A case for stronger scaffolding, Phys. Rev. ST Phys. Educ. Res. **7**, 010109 (2011).

[19] S. J. Osterlind and H. T. Everson, *Differential Item Functioning* (Sage Publications, Thousand Oaks, CA, 2009).

[20] R. Zwick, *A Review of ETS Differential Item Functioning Assessment Procedures* (Educational Testing Service, Princeton, NJ, 2012).

[21] K. Wilson, D. Low, M. Verdon, and A. Verdon, Differences in gender performance on competitive physics selection tests, Phys. Rev. Phys. Educ. Res. **12**, 020111 (2016).

[22] R. D. Dietz, R. H. Pearson, M. R. Semak, and C. W. Willis, Gender bias in the force concept inventory?, AIP Conf. Proc. **1413**, 171 (2012).

[23] L. J. Rennie and L. H. Parker, Equitable measurement of achievement in physics: High school students' responses to assessment tasks in different formats and contexts, J. Women Minorities Sci. Eng. **4**, 113 (1998).

[24] L. McCullough, Gender, context, and physics assessment, J. Int. Wom. Stud. **5**, 20 (2004).

[25] W. McBride *Teaching to Gender Differences: Boys Will be Boys and Girls Will be Girls* (World Books, Chicago, IL, 2009).

[26] D. Wood, J. S. Bruner, and G. Ross, The role of tutoring in problem solving, J. Child Psychiat. **17**, 89 (1976).

[27] T. C. Lin, Y. S. Hsu, S. S. Lin, M. L. Changlai, K. Y. Yang, and T. L. Lai, A review of empirical evidence on scaffolding for science education, Int. J. Sci. Math Educ. **10**, 437 (2012).

[28] C. Quintana, B. J. Reiser, E. A. Davis, J. Krajcik, E. Fretz, R. G. Duncan, E. Kyza, D. Edelson, and E. Soloway, A scaffolding design framework for software to support science inquiry, J. Learn. Sci. **13**, 337 (2004).

[29] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevPhysEducRes.13.020117 for all long-answer exam questions belonging to the low and high scaffolding groups.